The Application of Multi-objective Genetic K-means Clustering Algorithm to the Analyses of Cardiac Disease Diagnosis

Jenn-Long Liu*[1], Chien-Ting Chen[2]

[1]Department of Information Management, I-Shou University, Kaohsiung 84001, Taiwan

[2]Department of Information Engineering, I-Shou University, Kaohsiung 84001, Taiwan

*[1]jlliu@isu.edu.tw; [2]arthur@fotech.edu.tw

Abstract

In a case of misdiagnosing of disease, false negativity can cast a much larger impact than false positivity for both the doctor and the patient. Recently, many researchers have devoted to the study of using data mining on disease diagnosis. Current research shows by emphasizing the adoption of various clustering or classification techniques in cardiac disease diagnosis, one fails to lower the risk of misjudging a case as false negative. Therefore, this research focuses on the combination of multi-objective genetic algorithm and k-means clustering technique in data mining to analyze the data of patients suffering from cardiac disease, hoping to get higher prediction accuracy in the diagnosis and lower the risk of misjudging a case as false negative, the result of which needs to pay a high cost. These two algorithms are combined and named Multi-objective Genetic k-means Clustering Algorithm, termed MOGKCA in this work. The objectives of multi-objective optimization on cardiac disease analysis are: (1) to minimize the inaccuracy of classification (or maximize the prediction accuracy of classification), and (2) to minimize the number of false negativity (or minimize the cost of classification). The data used in this research are from UCI cardiac disease data set, Heart-staglog. From the single objective optimization case, the proposed MOGKCA can effectively elevate the classification rate. In the analysis of the two cases of data sets in their multi-objective optimization, we can get the best Pareto-Front through MOKGCA, which can be provided and help set the best point of analysis.

Keywords

K-means Clustering; Cardiac Disease Diagnosis;Multi-objective Genetic k-means Clustering Algorithm; False Negativity

Introduction

Among different cardio-vascular diseases, cardiac disease lists number 2 among the top ten causes of death domestically. In recent years, we can see the mature analysis of patient data from case data bases. In view of this, this work adopts a Multi-objective Genetic Algorithm (MOGA) with the combination of the clustering technology of k-means to analyze case data bases of cardiac disease, hoping to establish a model based upon multi-objective classification.

In data mining, the confusion matrix (Table 1) is generally used to evaluate the computing accuracy of data bases to achieve the classified accuracy (CA), which is *CA= (tn+tp)/ (tn+tp+fn+fp)*. Thus the inaccuracy is 1-CA. However, in clinical diagnosis, cost matrix still has to be counted in addition to the consideration of classification accuracy. This is because in diagnosing a medical case, the inaccuracy lies mainly in false positivity and negativity. The inaccuracy of false positivity means the ratio a patient is with no disease but is diagnosed contrarily. False negativity means the ratio of a patient being diagnosed as having the disease while in reality he or she does not. In practical medical diagnosis, the ratio of false negativity causes much more impact than that of false positivity on the doctor and patient. Accordingly, necessary cost resulting from misdiagnosis is analyzed to form a cost matrix as shown in Table 2. Table 2 shows the cost needed to be shouldered when an inaccuracy occurs. We can observe from the matrix that 5 units of cost will have to be paid once a healthy person is misdiagnosed with a disease. On the contrary, one unit of cost will have to be paid when a sick person is misdiagnosed without a disease. This shows the former has to pay 5 times as much as the later. Therefore, seeing Table 2, the cost from diagnosing mistakenly is 5×fn+fp. The higher the ratio of false negativity, the higher cost will have to be paid, but the result will be far from being satisfactory.

Judging from the above, in the analysis of medical disease, we not only have to focus on getting higher accuracy in prediction (high ratio of true positivity and negativity), we should try equally hard to reach a state of lower cost (low ratio of false negativity). Therefore, applying data mining technology to the analysis of disease, we have to simultaneously get a high accuracy and therefore pay a much lower cost by using the multi-objective optimization method.

TABLE 1 CONFUSION MATRIX OF CLASSIFICATION

Prediction Actuality	False	True
Negative	*tn*	*fp*
Positive	*fn*	*tp*

TABLE 2 COST MATRIX OF CLASSIFICATION

Prediction Actuality	False	True
Negative	*0*	*1*
Positive	*5*	*0*

In reality, many objectives of optimum can exist at the same time and their conflicts cannot be simplified. In this light, there is usually no one single best solution among these problems. Rather, many alternative solutions can combine to form one solution set. In general, when all objectives are considered and conditioned to match some constraints, this combined solutions form a Pareto-optimal front. Each solution on the Pareto-optimal front is named non-dominated solution. As Fig. 1 indicates, in a process where a single objective evolves, the predicted accuracy can be increase, so the cost will decrease, until it reaches its optimum marked with symbol "★". When multi-objective optimization reaches its optimum, the cost will be higher if a higher accuracy is to be demanded. In light of this, accuracy and cost can be conflicting. We have to know in predicting a disease; we have to set an acceptable minimum accuracy to start with in the pursuit of an optimum. On the premium that all conditions are met, the higher accuracy and lower rate of false negativity we can get, the better. The analyzer can find out the best decision making point from the Pareto-optimal front.

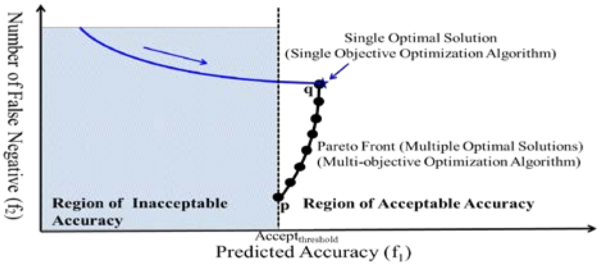

FIG. 1 DISTRIBUTION OF MULTI-OBJECTIVE SOLUTIONS

Literary Review

From the aforementioned, accurately diagnosing to see if a person is afflicted with cardiac disease can be of great help in maintaining people's health. Currently, some research works adopt mostly data-mining only, such as C5 decision making tree (Quinlan, 1996), CART (Breiman et al., 1984) and Back Propagation Networks (BPN) (Anthony and Bartlett, 2002). In similar studies in the literature, many have found that through supporting decision making systems designed by intelligent system can achieve lower cost in clinical tests (Palaniappan and Awang, 2008; Shantakumar and Kumaraswamy, 2009) proposed an intelligent prediction system by combining data mining with artificial neural network technology to predict the cardiac disease using the data set of Cleveland. First, they classified the data bases related to cardiac disease through k-means and yield necessary data, and they found the frequent pattern through Maximal Frequent Itemset Algorithm (MAFIA) and have concluded that the corresponding numerical results that influence cardiac disease should be: (1) blood pressure over 140/90 mmHg, (2) cholesterol over 240 mg/dl, (3) the maximum heart beat rate over 100 times every minute, (4) the abnormal readings on the electro-cardiogram, and (5) irregular angina pectoris. Then the multi-layer perceptron neural network with back-propagation can get rules of cardiac disease through data-mining. In addition, Parthiban and Subramanian (2008) proposed a coactive fuzzy inference system based method to predict cardiac disease. They also introduced a GA to find out the membership function and the best parameter that are required in CANFIS. They analyzed Cleveland cardiac disease data set and found that the system of genetic algorithm mixed with CANFIS can work well in cardiac disease prediction. Moreover, Rajeswari et al. (2011) mentioned the complex factors of forming coronary heart disease (CHD). The parameter of age, gender, cigarette, high blood pressure, high cholesterol, diet, long sitting, type 'A' blood, and family history all serve as significant factors in the forming of cardiac disease. They also employed GA to find 19 dangerous factors resulting in cardiac disease, and then gave them risk scores each. Srinivas et al. (2010) adopted One Dependency Augmented Naïve Bayes (ODANB) classifier and Naïve Creedal Classifier 2 (NCC2) in analyzing Heart-c, Heart-h and Heart-Statlog from UCI date sets. Their result showed 15 attributes resulting in cardiac disease. Yet the highest accuracy rate that they have achieved among the three was only 84,14%, which shows the ineffectiveness of only using data-mining way to achieve high accuracy. Kumaria and Godara (2011) adopted the methods RIPPER, Dicision Tree, ANNs and SVM in analyzing the Heart-h cardiac disease data collection from UCI. Their results showed that SVM has the highest accuracy rate of 84.12%, in which false negativity is also mentioned in accurately judging a case, and the lower its numerical results, the better. Das et al. (2009) adopted SAS Enterprise Miner 5.2 in analyzing Cleveland cardiac disease date set and got a high rate of 89.01% in classification and accuracy, 80.95% of sensitivity, and 95.91% of specificity. Anbarasi et al. (2010) pointed out in their cardiac disease diagnosis that the use of genetic algorithm lowers the scale of data and can get the best attributes in cardiac disease prediction, bettering the accuracy of decision making tree and the prediction of Bayesian Classifiers (Soni et al., 2011). Based upon studies, the combination of GA and data-mining can help find the attributes of some highly risk factors and prevent from having cardiac disease thereby, serving as a basis to help detect false negativity.

Basically GA belongs to one of the best heuristic searching. The reason for its use is to find the best solution through the evolution process from generation to generation. Generally, three functions of the genetic algorithm are: selection, crossover, and mutation. The GA can offer the best analysis, and by combining data-mining such as k-means, it can elevate the convergence of group k (Krishna and Murty, 1999). Roy and Sharma(Roy and Sharma, 2010) adopted GKMODE to analyze Cleveland cardiac disease set and show that the accuracy of classification reaches (130+110)/(130+110+34+29) =79.2%, false negativity is fn=34, and false positivity is fp=29. The accuracy can be heightened while false negativity can be further lowered. In addition, related literature also mentions the combination of the GA with k-means can become an effective tool to optimize classification.

Methodology

The importance of our research and related studies in this field mentioned above, our study focuses on: (1) organize the data related to cardiac disease and analyze some significant attributes, (2) combine data-mining clustering technology and MOGA to complete the evolving Multi-objective Genetic k-means Clustering Algorithm, termed MOGKCA, and (3) analyze cases of cardiac disease.

K-means Clustering Analysis

The research adopts k-means clustering, which is meant to divide the data on hand to k groups, each with a centroid. Clustering arises through the similarities of data, the similarity of which is judged through the measurement of distance or space. The generally employed equation is Manhattan distance. When there are n attributes, the Manhattan distance equation can be from $\vec{x}_i$ to $\vec{x}_c^j$ in (1), as seen below:

$$\sum_{i=1}^{n} \left| x_i - x_c^k \right| \tag{1}$$

Grouping is made according to the distance measured. This kind of grouping has two requirements: (1) at least a data point has to be included, and (2) every data point has to belong solely to one group. Combining the GA to provide the weight of every attribute, the research decides to adopt Manhattan distance. The distance from between the two points times the weight before grouping. The distance of similarity is reached as the equation shown below:

$$\sum_{i=1}^{n} w_i \times \left| x_i - x_c^k \right| \tag{2}$$

The data set here is divided into two groups (*k*=2), those with and without cardiac disease. The process of k-mean algorithm goes as follows:

(a). Choose k.

(b). Choose k initial centroids.

(c). Make use of similarity computing, and grouping data points to their nearest data points.

(d). Use these data points in the grouping centers, and recalculate the new data points.

(e). Check step (d) to see if the center points alter. If so,steps (c) and (d) will be repeated until the center points remain the same again. In this state, we reach the convergence of k-means algorithm.

Multi-objective Genetic Algorithm

The MOGA used in this study(Liu et al., 2010) includes: (a) put the individual solution in the group solutions according to the definition of dominate, and perform a non-dominated sorting to each individual (Deb et al., 2002), (b) calculate the crowded distance for each solution and sort them, (c) use the method of binary tournament selection operator, (d) apply a extend intermediate crossover, and (e) use a non-uniform mutation operator (Deb et al.; Liu et al.). By doing so, the effectiveness of the MOGA can be heightened. The optimization of multiple objectives can be formulated as a minimization of function $\vec{f}(\vec{x})$ with M objectives subject to p constraints denoted by function $g(\vec{x})$ as follows:

$$\text{Minimize } \vec{f}(\vec{x}) = \left\{ f_1, f_{2,} ..., f_M \right\}^T \qquad (3)$$
$$\text{Subject to } g_j(\vec{x}) \ge 0,\ j = 1, ..., p$$

In the minimizing process, if the situation meets the following two equations, we call it: " $\vec{x}_a$ dominates $\vec{x}_b$."

$$\forall i \in \{1, 2, ..., M\},\ f_i(\vec{x}_a) \le f_i(\vec{x}_b)$$
$$\exists j \in \{1, 2, ..., M\},\ f_j(\vec{x}_a) < f_j(\vec{x}_b) \qquad (4)$$

In the process of optimum, the dominated solutions will gradually recede into non-dominated ones. When $\vec{x}^*$ is found, so is the optimum, and we can obtain a Pareto-front.

The Combination of MOGA and K-means Clustering Algorithm

This study makes use of GA to find a best possible weight ($\vec{w}$), then combines the result with that similarity computing (Eq. (1)) in k-means algorithm. Then we put the misjudging value of false negativity yielded from k-means into the use of GA, thereby setting the dual objective function to get the best solution ($\vec{w}_{opt}$). The fitness functions we use here are: (1) minimal classification inaccuracy, and (2) minimal number of false negativity we get after minimizing classification, presented as follows:

$$f_1(\vec{w}) = Min\left(\sum_{i=1}^{n} \left| \left(C_{pred}\right)_i - \left(C_{actual}\right)_i \right| \right)$$
$$f_2(\vec{w}) = Min\left(\sum_{i=1}^{n} \left| fn \right| \right) \qquad (5)$$

Our search put together k-means algorithm and multi-objective genetic algorithm, renaming it Multi-objective Genetic k-means Clustering Algorithm (MOGKCA).

Discussion and Result

The cardiac disease data set Heart-staglog comes up which is from UCI (Blake and Merz, 1998). First, Heart-staglog has 13 attributes and one class attribute. The total number of the data set is 270. Two classifications were made from the class attribute (120 cases with cardiac disease, 150 cases without).Our results go as following: (1) the best optimum prediction in single objective and (2) the best optimum result in multiple objectives.In single objective prediction, $f_1(\vec{w})$ in (8) is adopted to get an optimal result, and the cardiac disease cases are from UC: Heart-staglog, which adopts (1)100% data as training and test sets and (2) 80% of data as training set, and 20% as test set. There are 13 attributes and one classified result in the collection of patient medical history, as shown in Table 3. Every case has undergone twenty times of independent runs. The result will be compared using confusion matrix and performance measurement parameters, which are presented as follows:

(i) Accuracy: $accuracy = \dfrac{tn + tp}{tn + fn + fp + tp}$

(ii) Sensitivity: $sensitivity(P) = \dfrac{tp}{fn + tp}$

(iii) Specificity: $specificity(R) = \dfrac{tn}{tn + fp}$

(iv) F-measure: $F - Measure = \dfrac{2 \times P \times R}{P + R}$

(v) Cost: $\cos t = fp + fn \times 5$

TABLE 3 ATTRIBUTES AND DOMAINS OF DATA SET OF CARDIAC DISEASE

No	Attributes	Domain
1	Age (age)	continuous
2	Sex (sex)	male/female
3	Chest (cp)	1, 2, 3, 4
4	Resting blood pressure (trestbps)	continuous
5	Serum cholesterol (chol)	continuous
6	Fasting blood sugar (fbs)	0, 1
7	Resting electrocardiographic results (restecg)	0, 1, 2
8	Maximum heart rate achieved (thalach)	continuous
9	Exercise induced angina (exang)	0, 1
10	Old peak	continuous
11	Slope (slope)	1, 2, 3
12	Number of major vessels (ca)	0,1,2,3
13	Thal (thal)	3, 6, 7
14	Class (class)	healthy or sick

Results of Single Objective Optimization

The single objective optimization is got through the adoption of $f_1(\vec{w})$ in (8) as the objective. Table 4 lists the contrast of confusion matrix. The average accuracy rate was 89.48% after conducting MOGKCA 20 times. The standard deviation was 0.4596%. The best accuracy rate of classification was 90.00%, while the cost and the number of false negativity were 95 and 17, respectively. On the contrary, the accuracy rate was 83.70% for the adoption of k-means clustering algorithm only, and had a cost of 168 with 13 cases of false negativity. From these contrast, we can see the proposed MOGKCA can highly increase the accuracy in classification while lowering the cost from misdiagnosis. Likewise, from Table 5, we can also see this same technology greatly improves what k-means clustering has achieved. Compared with Naïve Bayes and See5, GA k-means gets a result similar to that of

See 5.

TABLE 4 COMPARISON OF CONFUSION MATRIX FOR SINGLE OBJECTIVE OPTIMIZATION

reality \ prediction		healthy	sick
healthy	k-means	137	13
	MOGKCA	140	10
sick	k-means	31	89
	MOGKCA	17	103

TABLE 5 COMPARISON OF SINGLE-OBJECTIVE EFFECTIVENESS

parameter \ algorithm	Naïve Bayes	See5	k-means	MOGKCA
accuracy	86.29%	90.00%	83.70%	90.00%
sensitivity	83.33%	87.50%	74.17%	85.83%
specificity	88.67%	92.00%	91.33%	93.33%
F-measure	85.92%	89.69%	81.86%	89.43%
cost	117	87	168	95

Prediction of Multi-objective Optimization

The first goal of multi-objective optimization was to set a minimization of classification inaccuracy. The second one was to lower down the cases of false negativity. Restriction condition required that the accuracy of classification be more than 80%. Therefore, in this case, $f_1(\vec{w})$ and $f_2(\vec{w})$ in (8) were chosen to be the goal towards an optimum, thereby multi-objective analysis can be really started.

1) Case 1: The Total Number of Training Set Data is 270 (test set number equals that of training set)

All the 270 served to be the data of training set. Test set number was also 270. As the result shows, the best Pareto-front was shown in Fig. 2. Every point between "p" and "q" on the front was a Pareto-optimal solution. The solution of "p" possessed lower rate of false negativity, yet the accuracy was low. On the contrary, the solution of "q" had a higher rate of accuracy, yet the number of false negativity was also high. Table 6 lists the comparison of solutions "p" and "q" in confusion matrix. Average values and standard deviation after 20 times of classification were in Table 8. The solution of "q" had a higher rate of accuracy in classification but a lower rate of standard deviation. In addition, we can calculate and know the accuracy rates of optimal solutions of "p" and "q" from Table 7, which were 82.59% and 90.37% respectively. The numbers of false negativity in points "p" and "q" were 9 and 17 each. This showed multi-objective genetic k-means can effectively presents the best Pareto solution after multi-objective optimization. Table 8 is comparison of the effectiveness of optimal solutions at "p" and "q". The solution of "p" was higher sensitivity but with a lower cost. Yet that of point "q" had higher accuracy of classification, specificity, and F-measure. Therefore, by analyzing what we've got from Pareto-front, we can offer analyzers the best points to analyze. In addition, we can see from Table 9 that two attributes coming from the solutions of points "p" and "q" have remarkable influence by attributes "cp" and "ca". Yet they show least effect on the examination of the 9th attribute 'exang'.

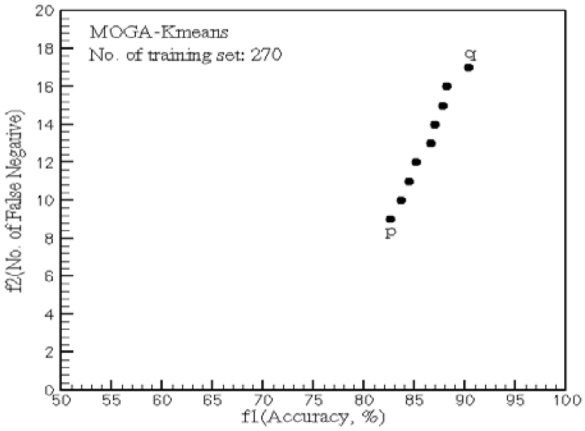

FIG. 2 DIAGRAM OF DISTRIBUTION OF MULTI-OBJECTIVE OPTIMUM SOLUTION IN CASE 1.

TABLE 6 COMPARISON OF SOLUTIONS "P" AND "Q" IN CONFUSION MATRIX FOR CASE 1

reality \ prediction		healthy	sick
healthy	solution at "p"	112	38
	solution at "q"	141	9
sick	solution at "p"	9	111
	solution at "q"	17	103

TABLE 7 AVERAGE VALUES AND STANDARD DEVIATION FOR CASE 1

item \ solution point	"p"	"q"
average	83.70%	89.46%
standard deviation	1.45818%	0.54275%

TABLE 8 COMPARISON OF THE EFFECTIVE OF OPTIMAL SOLUTIONS AT "P" AND "Q" FOR CASE 1

parameter \ solution point	"p"	"q"
accuracy	82.59%	90.37%
sensitivity	92.50%	85.83%
specificity	74.67%	94.00%
F-measure	82.63%	89.73%
cost	83	94
No. of false negativity	9	17

TABLE 9 WEIGHTS OF DIFFERENT ATTRIBUTES FOR CASE 1

attributes	weights at"p"	weights at "q"
w1 (age)	0.10590	0.67609*
w2 (sex)	0.39751	0.49020
w3 (cp)	0.83539*	0.80064*
w4 (trestbps)	0.15497	0.32295
w5 (chol)	0.10438	0.12337
w6 (fbs)	0.11371	0.05776
w7 (restecg)	0.41222	0.29321
w8 (thalach)	0.21988	0.06001
w9 (exang)	0.00696	0.00087
w10 (oldpeak)	0.61670*	0.46671
w11 (slope)	0.42300	0.46982
w12 (ca)	0.79464*	0.53327*
w13 (thal)	0.23767	0.19470

2) *Case 2: 80% of Data Sets (216 cases) as Training Set; the Rest (20%, 54cases) as Test Set*

In Case 2, 80% of data sets (216 cases) were training sets (randomly chosen). The remaining 20% (54cases) were test sets. The result of thePareto-Front was displayed in Fig. 4. The training set number was lower compared with Case 1, so the accuracy of training set classification dropped down a bit. In Fig. 3, the Pareto-optimal solutions were from points "p" to "q" in Case 2. The solutions on the front were fewer than Case 1. As Pareto-Front showed, the solution of point "p" came with a smaller number of false negativity. Yet the classification accuracy was lower. Solution "q" was more accurate in classification. Yet the number of false negativity was also larger.

Table 10 shows confusion matrix comparison between solutions "p" and 'q.' Table 11 shows the average classification accuracy and its standard deviation after 20 times of multi-objective genetic k-means calculation. The classification accuracy rates were 84.90% and 87.59%, and the standard deviation rates were 3.27495% and 1.76655%, respectively. The solution of point "q" had higher classification accuracy and lower standard deviation. The two solutions of "p"and"q"as applied to the test set were 81.48% and 90.74%. From Table 12, we came up with the best classification accuracy of points "p" and "q", which were 82.41% and 89.81% respectively. Table 12 lists the comparison of the effectiveness of the two solutions of"p"and"q". The solution of "p" had higher sensitivity and lower cost; whereas that of "q" had higher classification accuracy, specificity, and F-measure. The numbers of false negativity of "p" and "q" were 8 and 14. Still, the solutions "p" and "q" were remarkable in the attributes "cp" and "old peak", while it shows the least effect in attribute "exang" in the case.

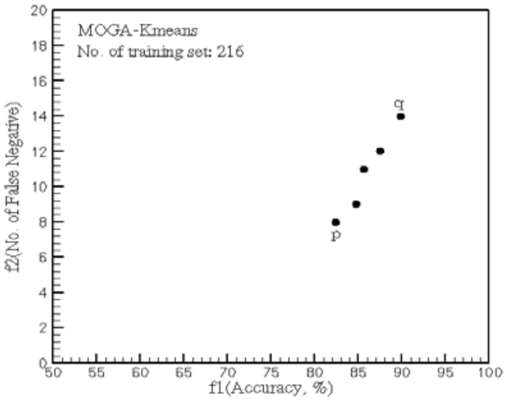

FIG. 3 MULTI-OBJECTIVE OPTIMAL SOLUTION DISTRIBUTION FOR CASE 2.

TABLE 10 COMPARISON OF CONFUSION MATRIX AT"P" AND "Q" FOR CASE 2

prediction / reality		healthy	sick
healthy	solution at "p"	95	30
	solution at "q"	117	8
sick	solution at "p"	8	83
	solution at "q"	14	77

TABLE 11 AVERAGE CLASSIFICATION VALUE AND STANDARD DEVIATION FOR CASE 2

solution point / item	"p"	"q"
average	84.90%	87.59%
standard deviation	3.27495%	1.76655%

TABLE 12 COMPARISON OF THE EFFECTIVENESS OF OPTIMAL SOLUTION AT "P" AND "Q" FOR CASE 2

solution point / parameter	"p"	"q"
accuracy	82.41%	89.81%
sensitivity	91.21%	84.62%
specificity	76.00%	93.60%
F-measure	82.91%	88.88%
cost	70	78
No. of false negativity	8	14

Conclusion

The study proposed Multi-objective Genetic k-means Clustering Algorithm to analyze the patient cases in cardiac disease data base and heightens the diagnostic accuracy, while lowering the cost due to false negativity. We adopted materials from UCI cardiac disease data set: Heart-staglog. From the single objective optimization case, we know the proposed MOGKCA can effectively elevate the classification rate. In the analysis of the two cases of data sets in their multi-objective optimization, we can get the best Pareto-Front through MOKGCA, which can be provided and help set the best point of analysis.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of Republic of China under project No. 103-2918-I-214-001.

REFERENCES

Anbarasi, M., Anupriya, E., and Iyengar, N.CH.S.N."Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm."International Journal of Engineering Science and Technology, Vol. 2, No. 10, 5370-5376, 2010.

Anthony M., and Bartlett, P. L. Neural Network Learning: Theoretical Foundations. Cambridge Uni. Press, 2002.

Blake, C.L., and Merz, C.J. UCI Repository of Machine Learning Databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

Breiman, L., Friedman, J., Olshen, R., and Stone, C.Classification and Regression Trees, Wadsworth International Group, 1984.

Das R., Turkoglu I., and Sengur, A. "Effective Diagnosis of Heart Disease through Neural Networks Ensembles." Expert Systems with Applications, Vol. 36, Issue 4, 7675-7680, May 2009.

Deb, K., Pratap, A., Agarwal, A., and Meyarivan, T. "A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II." IEEE Trans. on Evolutionary Computation, Vol. 6, No. 2, 182-197, 2002.

Krishna K., and Murty, M.N. "Genetic K-Means Algorithm." IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 29, No. 3, 433-439, June 1999.

Kumari, M., and Godara, S. "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction." International Journal of Computer Science and Technology, Vol. 2, Issue 2, 304-308, June 2011.

Liu, J.L., Chao, C.W., and Chen, C.M. "Optimizing Mobile Base Station Placement Using an Enhanced Multi-objective Genetic Algorithm." International Journal of Business Intelligence and Data Mining, Vol. 5, No. 1, 2010, 19-42.

Palaniappan, S., and Awang, R. "Intelligent Heart Disease Prediction System Using Data Mining Techniques." International Journal of Computer Science and Network Security, Vol. 8 No. 8, 343-350, August 2008.

Parthiban, L., and Subramanian, R. "Intelligent Heart Disease Prediction System Using CANFIS and Genetic Algorithm." International Journal of and Life Sciences, Vol. 3, No. 3, 157-160, 2008.

Quinlan, J. R. "Improved Use of Continuous Attributes in C4.5." Journal of Artificial Intelligence Approach, Vol. 4, 77-90, 1996.

Rajeswari, K., Vaithiyanathan, Dr. V., and Amirtharaj, Dr. P. "Prediction of Risk Score for Heart Disease in India Using Machine Intelligence." 2011 International Conference on Information and Network Technology, IPCSIT, Vol. 4, 2011, IACSIT Press, Singapore.

Roy D.K., and Sharma, L.K. "Genetic K-means Clustering Algorithm for Mixed Numeric and Categorical Data Sets." International Journal of artificial Intelligence & Applications, Vol. 1, No. 2, 23-28, April, 2010.

Shantakumar B.P., and Kumaraswamy, Y.S. "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network." European Journal of Sci. Research, Vol. 31 No. 4, 642-656, 2009.

Soni, J., Ansari, U., Sharma, D., and Soni, S. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction." International Journal of Computer Applications, Vol. 17, No. 8, 43-48, March 2011.

Srinivas, K., Rani, B.K., and Govrdhan, Dr. A. "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks." International Journal on Computer Science and Engineering, Vol. 2, No. 2, 250-255, 2010.